**Доверенный искусственный интеллект**

# Проблемы доверия искусственному интеллекту

А.Е. Самотуга, П.С. Ложников

Конспект лекции

# Проблемы доверия искусственному интеллекту

## Доверие к системам искусственного интеллекта

Доверие к системам искусственного интеллекта является важнейшим условием, определяющим возможность применения этих систем при решении ответственных задач обработки данных. Примерами таких задач являются поддержка принятия врачебных решений, беспилотное управление транспортными средствами и некоторые другие, ошибки при решении которых могут привести к тяжким последствиям, связанным с угрозой жизни и здоровью людей, серьезным экономическим и экологическим ущербом.

## Уровни доверия

На текущий момент принято выделять три уровня доверия, то есть по одному для каждого из уровней архитектуры ИИ (физический уровень, инфраструктурный уровень, прикладной уровень).

Согласно ГОСТ Р 59276-2020 для каждого из уровней доверия определяются свои существенные характеристики.

Доверие на физическом уровне, как правило, основывается на комбинации требований к надежности, безопасности и функциональности, поскольку существенные характеристики, требования к которым устанавливаются, основаны в данном случае на физических измерениях или тестах. Например, успешно пройденный технический контроль ТС делает ТС и его системы заслуживающими доверия. В этом контексте уровень доверия может быть определен посредством выполнения проверочных мероприятий. Кроме того, некоторые процессы, такие как калибровка датчиков, могут гарантировать правильность измерений и, следовательно, корректность полученных данных.

Доверие на инфраструктурном уровне заключается, как правило, в выполнении требований безопасности ИТ-инфраструктуры, таких как контроль доступа и другие меры для поддержания целостности и доступности системы ИИ, а также обеспечения конфиденциальности обрабатываемых данных.

Доверие на прикладном уровне системы ИИ требует, среди прочего, подтверждения надежности и безопасности ПО. Разработка ПО предполагает реализацию процессов его верификации и валидации. Обеспечение доверия к системам ИИ включает подходы, применимые к обычным информационным системам, но не ограничивается ими. Так, например, для систем ИИ, основанных на машинном обучении, способность вызывать доверие подразумевает также непредвзятость (объективность) функционирования системы, что соответствует отсутствию необоснованного смещения формируемых оценок.

1. **Компоненты и свойства доверенного ИИ**

Доверие к системе ИИ достигается в том случае, если выполняются требования к представительному набору существенных характеристик систем ИИ на всех трех уровнях. При этом используются способы обеспечения доверия на всех стадиях ЖЦ системы ИИ.

Очевидно, что доверие к системам ИИ — важнейшее условие их применения при решении ответственных задач обработки данных. Основные компоненты доверенного ИИ - проверяемость (объяснимость), управляемость, стабильность, отказоустойчивость, безопасность и робастность.

Проверяемость (прозрачность, объяснимость) ИИ. Проверяемость подразумевает, что ИАС не является черным ящиком. Объяснимость и прозрачность ИИ подразумевает среди прочего четкое и внятное объяснение для пользователя, с каким типом интерфейса он общается в данный момент: с человеком, с человеком, которому ассистирует ИИ, или только с ИИ.

Управляемость. Интеллектуальные системы должны быть управляемы человеком. Он должен иметь возможность применять весь спектр воздействий, от выключения или включения системы до управления в ручном режиме.

Стабильность, безопасность, отказоустойчивость. Для технической устойчивости важны надежность инфраструктуры (в первую очередь доступность интернета), которая требует постоянной поддержки, и наличие альтернативных аналоговых путей.

Робастность, то есть устойчивость к внешним воздействиям. Под такими воздействиями имеются в виду ситуации, когда, например, в выборку из миллиона фото, на которой работает алгоритм распознавания лиц, добавили еще одну фотографию. Если алгоритм «ломается», перестает правильно распознавать лица, это говорит об отсутствии робастности.

Защита персональных данных. Разработчик должен позаботиться о том, чтобы ПДн нельзя было извлечь из системы.

Право отказаться от использования ИИ. С этической точки зрения не следует принуждать людей к использованию цифровых платформ, даже если кому-то такой выбор кажется безусловным благом.

Равный доступ. В государстве не должно быть преференций для доступа к ИАС у отдельных категорий граждан, выделенных по социальному, экономическому, политическому признаку.

Преодоление дискриминации. В ИИ-системах может быть (непреднамеренно) встроена дискриминация людей по расовым, этническим, социологическим и прочим признакам; надо стремиться тому, чтобы подобных ситуаций не было.

1. **Аппаратные платформы для доверенного ИИ**

На сегодня есть разные по реализации систем доверенного ИИ. Выделяют программные, аппаратные и облачные решения. Полностью программные решения позволяют использовать произвольные вычислительные устройства, главное условие при этом – это их соответствие требованиям к вычислительным ресурсам. В чистом виде программные решения для ИИ встречаются очень редко, поэтому имеет смысл подробно остановиться на платформах ИИ и аппаратных платформах, также упомянуть облачные платформы. Разработками аппаратных платформ для ИИ занимаются многие в т.ч. известные корпорации. Так на рынке есть встраиваемые решения, на основе которых и создаются аппаратные платформы:

1) Платформа Jetson, включающая несколько устройств - Jetson Tx1|Tx2 и другие.

2) Raspberry Pi – состоит из нескольких поколений устройств, достаточная производительность появилась с RPi4.

3) Google Coral – включает в себя один чип и два девайса — Dev Board и USB Accelerator.

4) Intel Movidius – самые доступные варианты - Movidius 1|Movidius 2.

5) Gyrfalcone – разработано два поколения 2801, 2803.

Частично сюда можно отнести:

1) Intel процессоры

2) Nvidia мобильные GPU

3) Jetson AGX Xavier

4) Смартфоны

5) GAP8

6) Grove AI HAT.

Помимо упомянутых устройств, следует также отметить что оборудование для работы с ИИ-алгоритмами уже выпускают Cisco, Nvidia, Intel и др.

Google анонсировала процессор для машинного обучения — Tensor Processing Units — еще в 2016 году. Сегодня воспользоваться вычислениями на TPU могут клиенты Google Clouds. Amazon Web Services в прошлом году представила микросхему AWS Inferentia, разработанную специально под задачи машинного обучения.

1. **Понятие рисков ИИ. Уязвимости ИИ и новые угрозы безопасности. Дрейф модели и предвзятость ИИ. Непредсказуемость ИИ. Непрозрачность ИИ**

Несмотря на все преимущества, искусственный интеллект также является источником новых рисков, которыми надо управлять. Поэтому важно определить риски, которые относятся к каждой отдельной программе ИИ и каждому бизнес-подразделению, использующему данную технологию.

Внедрение и применение систем ИИ, как и других новейших технологий, неотрывно связано с появлением целого набора новых рисков. В этой связи, появляется необходимость в разработки специального подхода по оценки таких рисков.

Прежде всего. следует определиться с тем, что же входит в риски, специфические для ИИ. Как известно, для построения любой модели ИИ, нужны данные. При этом данные (в том числе, могут быть персональными), которые применяются на этапе создания модели ИИ. являются очень чувствительными. Они во многом определяют существенные характеристики построенной модели, ее качество, точность полученных результатов и поведение модели с течением времени.

К специфическим угрозам ИИ относятся возникновение аномалий данных (утечек данных, отравления данных, смещения моделей и другие), угрозы, связанные с использованием состязательных атак.

С этими угрозами связан ряд рисков:

* Необъективность, вызванная предвзятыми наборами данных и алгоритмами;
* Новые угрозы безопасности для систем ИИ (например, состязательные атаки).
* Плохая спецификация, плохая реализация или неудовлетворительные верификация и валидация могут привести к ненадежной работе систем ИИ.
* Нехватка прозрачности и прочие риски.
* Недостаток эксплуатационной надежности, например, при неисправности оборудования, вследствие сложности и непрозрачности систем.
* Непредсказуемость ИИ может быть связана с недостатком прозрачности, понятности, управляемости системы. Проблемами на этапе интеграции с другими системами.

В целом, управление рисков для ИИ состоит из тех же этапов, что и в других областях. Все начинается с идентификации рисков, после чего выполняют анализ рисков, оценку риска, обработку риска с выбором вариантов обработки.

1. **Принципы ответственного использования**

Помимо этого, для того, чтобы можно было говорить о доверии к ИИ системам, важно соблюдать определенным принципы ответственного использования:

* Правомерность: прикладные технологии ИИ будут разрабатываться и применяться в соответствии с внутренним законодательством и международным правом, включая международное гуманитарное право и право в области прав человека, как целесообразно.

• Ответственность и подотчетность: прикладные технологии ИИ будут разрабатываться и применяться с надлежащей степенью ответственности; в целях обеспечения подотчетности устанавливается четкая ответственность человека.

• Объясняемость и отслеживаемость: прикладные технологии ИИ будут должным образом понятны и прозрачны, в частности благодаря использованию методологий, источников и процедур обзора. Это предусматривает механизмы проверки, оценки и подтверждения соответствия.

• Надежность: будет четко определено, в каких случаях применяются прикладные технологии ИИ. Технологическая и физическая безопасность, а также эксплуатационная надежность этих средств будут подвергаться проверке и гарантии качества, согласно заданным сценариям использования на протяжении всего жизненного цикла, в том числе в установленном порядке сертификации.

• Управляемость: прикладные технологии ИИ будут разрабатываться и применяться в соответствии с предусмотренными функциями и при этом будет обеспечено: надлежащее взаимодействие человека с машиной; способность выявлять непреднамеренные последствия и избегать их; способность предпринимать шаги, такие как отключение систем в тех случаях, когда эти системы демонстрируют непредусмотренное поведение.

• Снижение предвзятости: будут предприниматься активные шаги по сведению к минимуму непреднамеренной предвзятости при разработке и применении прикладных технологий ИИ, а также в массивах данных.

1. **Безопасность на объектах критической информационной инфраструктуры**

В случае использования систем ИИ на объектах критической инфраструктуры, для обеспечения безопасности таких объектов важно соблюдать требования законодательства в т.ч. ФЗ О безопасности критической информационной инфраструктуры Российской Федерации, применять общие рекомендации и так далее.

Применение систем ИИ может быть более или менее критичным (например, траектория движения автономного транспортного средства VS траектория робота-уборщика). Могут быть определены различные уровни критичности приложений. Критичность может коррелировать с уровнем автоматизации системы.

С юридической точки зрения критичность будет соответствовать воздействию/потенциальному воздействию на отдельных лиц. В зависимости от критичности систем в системах ИИ для принятия решений будут реализованы одна или несколько контрольных точек. В зависимости от критичности приложения можно сопоставить уровень критичности и допустимый уровень надежности.

Для систем ИИ потребность в безопасности может быть связана с:

* Защитой клиента и конечного пользователя ИИ. В случае возможного негативного влияния со стороны системы;
* Защитой общества - то есть тех, кто не является пользователем, но на кого решения системы ИИ может оказать влияние.
* Безопасность взаимодействия с физическими системами с использованием технологии ИИ зависит от выявления всех соответствующих уязвимостей системы ИИ и от реализации всех соответствующих мер для обеспечения предсказуемого поведения системы ИИ.

1. **Проблемы, связанные со спецификацией, внедрением и использованием систем ИИ**

Спецификации гарантируют, что поведение системы ИИ соответствует истинным намерениям оператор. Проблема спецификации возникает, когда есть несоответствие между идеальной спецификацией и выявленной спецификацией, то есть когда система ИИ не делает то, что мы от неё хотим.

Проблема внедрения ИИ вызвана рядом препятствий:

* Отсутствие нормальных релевантных данных, которые не содержат ложных предпосылок и не нарушают конфиденциальность.
* Второе препятствие – это бизнес-процессы в компаниях.
* Ключевое препятствие – нехватка навыков для внедрения ИИ.
* Достаточно существенная стоимость решений на основе ИИ.
* Зачастую сложность в оценке эффективности от внедрения и отсутствие четких целей внедрения.

1. **Жизненный цикл ИИ. Тестирование ИИ. Развертывание модели**

Жизненный цикл систем искусственного интеллекта сходен с жизненным циклом другого программного обеспечения и включает этапы и критерии перехода между ними. Начинается с этапа «Разработка идеи и концепции системы», заканчивается этапом «Снятие системы с эксплуатации», включает в себя 13 этапов.

Разработка системы ИИ предполагает процесс генерации моделей ИИ, их последующего тестирования, сравнения, выбора лучшей по какому-либо параметру и ее развертывание.

1. **Решения MLOps, ModelOps**

Есть два типа систем, который сейчас наиболее распространены для создания моделей, их развертывания и дальнейшего мониторинга - MLOps, ModelOps.

Важно понимать различия в таких продуктах. Оно кроется как в том, для кого они предназначены, так и направления их использования. MLOps – используются аналитиками данных для проведения экспериментов, поиска лучших вариантов моделей ИИ, чаще используется для проведения исследований небольшими научными коллективами, ModelOps – системы направлены на решения конкретных задач фирмы, для изменения ее бизнес процессов, конечный результат важен для исполнительных директоров, так как приносит реальную пользу от внедрения модели, например, ускоряя этапы бизнес-процесса, снижая издержки, повышая точность решений. ModelOps применяется целыми подразделениями по анализу данных.

**Список источников**

1. 14:00-17:00. ISO/IEC TR 24028:2020 [Электронный ресурс]. URL: https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/07/76/77608.html (дата обращения: 12.04.2022).